GIS-based Data Synthesis and Visualization

Duccio Rocchini[1,*], Carol X. Garzon-Lopez[2], A. Marcia Barbosa[3], Luca Delucchi[1], Jonathan E. Olandi[1], Matteo Marcantonio[1,4], Lucy Bastin[5,6], and Martin Wegmann[7]

[1]Fondazione Edmund Mach, Research and Innovation Centre, Department of Biodiversity and Molecular Ecology, Via E. Mach 1, 38010 S. Michele allAdige (TN), Italy
[2]UR "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN, FRE 3498 CNRS), 9 Université de Picardie Jules Verne, 1 rue des Louvels, FR-80037 Amiens Cedex 1, France
[3]Centro de Investigacao em Biodiversidade e Recursos Geneticos (CIBIO), InBIO Research Network in Biodiversity and Evolutionary Biology, University of Evora, 7004-516 Evora, Portugal
[4]Geoinformation in Environmental Planning Lab, Technische Universitt Berlin, Germany
[5]School of Computer Science, Aston University, UK
[6]Currently on secondment to Joint Research Centre of the European Commission
[7]University of Würzburg, Department of Remote Sensing, Oswald-Külpe Weg 86, 97074 Würzburg, Germany
[*]corresponding author: ducciorocchini@gmail.com, duccio.rocchini@fmach.it

July 15, 2016

Contents

List of Figures

1 Introduction: ecological informatics applied to data synthesis and visualization

In spatial ecology synthesizing and properly visualizing data in 2D systems is a key issue when aiming at explaining spatial patterns by spatial processes.

This has been demonstrated in a number of ecological and geographical studies, dealing with different scientific aims (see Rocchini et al. (2016) and literature therein).

Increasing availability of open ecological data through networks like the Global Biodiversity Information Facility (GBIF, http://www.gbif.org) or the DataONE (Data Observation Network for Earth), on the one hand, and remote sensing data (e.g. http://remotesensing.usgs.gov/) on the other, makes it necessary to promote methods for data synthesis and visualization.

In this book chapter we will deal with synthesis and visualization related to the following ecological issues: i) synthesizing species distribution models relying on virtual species, ii) visualizing spatial uncertainty in species distribution based on cartograms, iii) fuzzy methods to synthesize species distribution uncertainty, iv) remote sensing data synthesis by exploratory analysis and re-plotting data in new systems, iv) measuring and visualizing ecological diversity from space based on generalized entropy, v) neutral landscape for testing ecological theories.

We will make use of examples from the Free and Open Source Software GRASS GIS (Neteler et al., 2012) and R (R Development Core Team, 2016).

2 Synthesizing species distributions by virtual species

Virtual species represent a powerful approach to build species distributions based on known ecological parameters shaping species spread. In this book chapter, the habitat suitability map presented here was created using the package "virtualspecies" in the R software. Two bioclimatic variables were used as proxies of habitat suitability (Figure 1). The bioclimatic variables selected are: annual temperature (Bio1) and annual precipitation, obtained from the bioclim dataset at 1 km spatial resolution. The resulting map shows a species with a wide niche distributed at the extent of Europe (Figure 1).

3 Cartograms to synthesize and visualize sampling effort bias

In ecology, a number of studies have dealt with the prediction of species distribution and diversity over space and its changes over time based on a set of environmental predictors related to environmental variability, productivity, spatial constraints and climate drivers.

Species distribution models have been acknowledged as the most powerful methods to map the spread of plant and animal species. The basic approach used to create maps based on predictors is to rely on linear models to create gridded landscapes of potential distribution of species based on point or polygon local data. In most cases, the output is a density function in two dimensions representing the distribution Sx of the x species. In general boundaries are sharply defined based on thresholds of predictors/factors (e.g. when mainly based on land cover, see also Comber et al. (2013)) or continuous, if based on the continuous variability of predictors (e.g. the continuous variability of temperature).

Uncertainty in such models mainly derives from input data pseudoabsences (Foody, 2011) as well as from models bias, i.e. the error deriving from the selected model (GAM, GLM, Maximum entropy models, etc,). Hence the visualization of uncertainty in two dimensions is strongly suggested although it is disregarded in most cases. However, its importance is apparent (Comber et al., 2012). In fact areas with a high or low probability of distribution of a

species might be also in relation with a high or low error rate.

Concerning bias related to sampling effort, we will rely on one of the mostly used datasets in biodiversity studies at large spatial extents, namely the GBIF dataset.

GBIF data comprises a huge range of species occurrence observations collected with a wide variety of sampling approaches. It spans from well established plot censuses to direct observations collected during field trips. Consequently, some of the data points are at the center of censused grids (each point comprises the species located at a specific-size quadrant) or correspond to single observations of one (or more) individuals of the same species. These differences also depend on the methodologies used to observe/record occurrences per taxon. Plots, and plots within transects, are common practice in vegetation censuses, while transects, point counts and live traps are preferred in the case of animals.

Moreover, the variation in factors such as per country biodiversity monitoring schemes, funding schemes, focal ecosystems, accessibility to remote areas, among others, add another source of variation, especially at multinational scales (Barbosa et al., 2013).

Undoubtedly, all those sources of variation result in a non homogeneous sampling that has important consequences not only on the development of accurate species distribution maps but, more importantly, on the conservation and management decisions focused on such a distribution of biodiversity (Rocchini et al., 2011). In this book chapter we synthesize spatial uncertainty in the sampling effort of the GBIF data, by explicitly taking into account potential area effects of European countries.

In this study we aim at quantifying and mapping the uncertainty derived from the variation in observations due to differences in sampling efforts. In particular the use of cartograms is proposed, in which the shape of objects is directly related to a certain property, in our case to uncertainty. Cartograms build on the standard treatment of diffusion, in which the current density is given by:

$$J = v(r,t)p(r,t) \tag{1}$$

where v(r, t) and p(r, t) are the velocity and density at position r and time t. Refer to Gastner and Newman (2004) for additional information.

Cartograms facilitate the visualization of spatial uncertainty in he results by changing the size of the polygons based on the den- sity of information contained (number of observations, variation, etc). For example, using this strategy, the spatial distribution of a species (e.g. *Fagus sylvatica*) can be

represented in a coloured grid in which the colour represents the abundance of the species and the distortion of the shape of each grid cell might represent the sampling bias, i.e. more distorted cells have been oversampled with respect to the others (Figure 2).

4 Fuzzy methods to synthesize species distribution uncertainty

Beside sampling bias, shown in the previous section, taxonomic bias, related to thematic (semantic) accuracy, might occur when different operators / scientists deal with the association of each individual to a certain species / class / taxon. Fuzzy set theory should aid in maintaining uncertainty information related to each species (hereafter also generally related to class as in fuzzy set theory). The concept of fuzzy sets was first introduced by Zadeh (1965); thus, fuzzy set based approaches have been widely used in ecology dating back to 1980s.
The principle behind fuzzy set theory is that the situation of one class being exactly right and all other classes being equally and exactly wrong often does not exist. Conversely, there is a gradual change from membership to non-membership Gopal and Woodcock (1994).
A fuzzy set is defined as follows: let U denote a universe of entities u, the fuzzy set F turns out to be:

$$F = (u, \mu(u))|u \in U \tag{2}$$

where the membership function associates for each entity u the degree of membership into the set F.
The degree of membership $\mu(u)$ ranges in the interval [0,1], i.e. the real range between 0 and 1.
Hence, fuzzy sets might represent a good starting point for continuously mapping species, by relying on each species as:

$$Fi = (u, \mu_i(u))|u \in U \tag{3}$$

$$Fj = (u, \mu_j(u))|u \in U \tag{4}$$

In this case, for each species i and j a map is derived based on e.g. fuzzy training data taken in the field (probability of each individual to belong to a certain species) representing species probability of occurrence. In this case, according to Boggs (1949) uncertainty is explicit in the sense that a probability of occurrence of each sampled individual to each species is mapped

instead of a crisp set considering that species are exhaustively determined, with a 100% accuracy.

One major assumption leads to consider fuzzy sets as a powerful tool for maintaining uncertainty information when aiming at mapping and analysing species or in general taxa distribution patterns, i.e. the gradual and continuous probability of correct determination of a certain species rather than considering a complete accuracy in the determination process. A fuzzy determination of a species might be derived, as an example, as the probability of correct determination given different operators / scientists. Figure 3 represents an example for the foraminifera species *Keratella quadrata*. A map of presence of the species worldwide (per country / region) is shown together with the probability (as inverse distance) of occurrence of each individual with the species / group. The analysis was performed relying on the *fuzzySym* package (Barbosa , 2015) for the R software.

5 Remote Sensing data synthesis

5.1 Synthesizing remote sensing data by exploratory data analysis

In some cases, RS data are correlated to each other; as an example, a high reflectance in a certain region of the electromagnetic spectrum might be related to that in another one. In other cases, indices derived from RS data are implicitly correlated. This is the case when calculating texture measures.

5.1.1 Exploratory correlation of RS bands by hexagon binning

Hexagon binning is a powerful technique for synthesizing geographical data, especially those based on huge 2D matrices.

An example is provided starting from two Landsat images freely available in the North Carolina dataset of GRASS GIS (https://grass.osgeo.org/download/sample-data/). Hexagon binning (R package "hexbin") clumps into hexagons point clouds once matrices are imported to R by using the package *rgrass7* (Figure 4). In this case the typical shape of Landsat NIR infrared versus Landsat red is achieved. Contrary to normal plots, hexagon binning allows to also show the amount of points per each value in the point cloud.

5.1.2 Exploratory correlation among several layers: the ezample of texture measures

Texture measures allow to investigate the amount of variability in a neighbourhood. This has a number of crucial ecological repercussions, especially in biodiversity studies in which local spatial heterogeneity is used as a proxy of species diversity (Rocchini et al., 2016).

In most cases texture measures are implicitly correlated. Showing such correlation is important to synthesize the texture system and avoid redundant information.

We propose an example using GRASS GIS which allows calculating the texture measures in a neighborhood of pixels described in the benchmark paper by Haralick et al. (1973): i) the angular second moment, as a measure of local homogeneity; ii) the contrast, a gray-level variation with respect to neighbor pixels; iii) the correlation, a linear dependency value; iv) the variance in the neighboring moving window (see also `r.neighbors`); v) the entropy, an index of randomness; vi) the sum average; vii) the sum entropy; viii) the sum variance; ix) the difference in variance; x) the difference in entropy; xi) the inverse distance moment, i.e. the inverse of the previously described contrast measure; xii) the maximal correlation coefficient. We refer to Haralick et al. (1973) for a detailed description of all the measures. Figure 5 presents two of the aforementioned maps generated from a Landsat ETM+ image: entropy and variance.

Further, to show the amount of correlation of such measures R can allow building a powerful graphical matrix based on correlation coefficients based on the package *corrplot*, once data are imported from GRASS GIS in R by the `rgrass7` package. Figure 6, a straightforward correlation matrix allows to synthesize the amount of correlation among texture measures in a graphical manner.

Since most of the measures are strongly correlated, when modelling ecosystem complexity, texture measures correlation should be first synthesized by a graphical output and texture measures should be used with care since, by their very nature, they are expected to be correlated with each other.

5.2 Fourier transforms

Remote sensing data are a powerful input for studying landscape transformations in space and time. In some cases, such transformations cannot be inspected in the normal space but a transform is needed to highlight such changes.

The use of transforms in frequency spaces to measure variation in a sig-

nal has long been acknowledged. While methods exist based on orthonormal series (e.g. rectangular decomposition of waves, Walsh (1923), the mostly used method rely on continuous waves, mainly based on the Fourier transforms (Fourier, 1822).

When seeking for a method to detect landscape change based on continuous instead of classified information, one should rely on a (continuous) function which i) does not require a-priori field information nor ii) a specific model based on the data being used. In this view, Fourier transforms (Fourier, 1822) may represent the best algorithmic solution.
Let f(x) be a continuous function described into a spatial domain. Based on the Fourier theorem (Fourier, 1822) every f(x) can be transformed into a continuum of sinusoidal functions of varying frequency, as:

$$F(\omega) = \int_{-\infty}^{\infty} f(x) \mathrm{e}^{-2\pi i \omega} \, \mathrm{d}x \tag{5}$$

where ω = frequency, also known as radian frequency since it is expressed in radians per spatial units. In mathematical notation for discrete Fourier transforms f(x)F(ω).

Extending Eq. 5 to two dimensions implies considering a two-dimensional function f(x,y), e.g. a raster matrix. Its Fourier transform turns out to be:

$$F(\omega, \nu) = \int \int_{-\infty}^{\infty} f(x, y) \mathrm{e}^{-2\pi i (\omega x + \nu y)} \, \mathrm{d}x, \mathrm{d}y \tag{6}$$

where ω,ν= frequency coordinates.
In the Fourier space, high frequency values (high heterogeneity) are at the border of the image while low frequency values (high homogeneity) are at the center. Hence the higher the value of pixels at the border, the higher the heterogeneity / complexity of the whole image (Figure 8).

6 Synthesizing diversity measurements from space: the case of generalized entropy

From a practical point of view, distinct diversity measures are aimed to summarize a large multivariate data set into one single value based on distinct objectives and approaches. Therefore, as this operation will always result in a loss of information, it is generally understood that there is no ideal summary statistics capable of unequivocally synthesizing all aspects of diversity (Ricotta, 2005).

In this view, Rényi (1970) proposed a generalized entropy, $H_\alpha = \frac{1}{1-\alpha} \times ln \sum p^\alpha$ which is extremely flexible and powerful since many popular diversity

indices are simply special cases of H_α. As an example, for $\alpha = 0$, $H_0 = ln(N)$ namely the logarithm of richness (N = number of DN values), i.e. the maximum Shannon entropy index (Hmax) which is used as the denominator of the Pielou index, while for $\alpha = 2$, $H_2 = \ln(1/D)$ where D is the Simpson Dominance index. For $\alpha = 1$ the Rényi entropy is defined in the limiting sense using l'Hospital's rule of calculus, and H_1 = Shannon's entropy H. Rényi's framework offers a continuum of possible diversity measures, which differ in their sensitivity to rare and abundant DNs, becoming increasingly regulated by the commonest DNs when increasing the values of α. In this view, changing α can be considered as a scaling operation that takes place not in the real but in the data space. That is why Rényi generalized entropy has been referred to as a continuum of diversity measures (Ricotta et al., 2003).

Changing the parameter α will change the behaviour of the formula generating different maps of diversity as represented in Figure 7, representing a continuum of diversity values over space instead of single measures. Increasing alpha values the Rényi diversity index will weight more differences in relative abundance instead of simple richness.

7 Neutral landscapes

Patterns in the field can be correlated to random patterns by calculating as an example the deviation from random expectations in two dimensions (Hanspach et al., 2011). To accomplish this goal, different kinds of lattice surfaces can be generated, including: completely random surfaces, gaussian distribution, fractal surfaces with a predefined fractal dimension.

This helps to make a comparison of real patterns found in landscape ecology with neutral landscape to synthesize if the real patterns show a significant deviation from random (neutral) expectations.

Random surfaces can be generated as in Figure 9, against which a Landsat image might be tested, to find as an example clumped parts of a Landsat image which significantly deviate from random expectations over space. Otherwise, a more sophisticated but still straightforward neutral model is represented by a gaussian surface, which should not be graphically different from a random surafce, but in this case values have a normal distribution in two dimensions, and the mean and the standard deviation can be defined a-priori.

A third example is represented by fractal surfaces Mandelbrot and Blumen (1989). Surfaces with a given fractal dimension from 2 to 3 might represent severe differences in their roughness / complexity (Imre et al., 2011)

(Figure 10). They might be used to test for the copmplexity of real patterns against such lattice images.

References

Barbosa, A.M. (2015). fuzzySim: applying fuzzy logic to binary similarity indices in ecology. Methods in Ecology and Evolution, 6: 853-858.

Barbosa, A.M., Pautasso, M., Figueiredo, D. (2013). Species- people correlations and the need to account for survey effort in biodiversity analyses. Diversity and Distributions, 19: 1188-1197.

Boggs, S. (1949). An atlas of ignorance: A needed stimulus to honest thinking and hard work. Proceedings of the American Philosophical Society, 93: 253-258.

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., Foody, G.M. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. International Journal of Applied Earth Observation and Geoinformation, 23: 37-48.

Comber, A., Fisher, P., Brunsdon, C., Khmag, A. (2012). Spatial analysis of remote sensing image classification accuracy. Remote Sensing of Environment 127: 237-246.

Foody, G.M. (2011). Impacts of imperfect reference data on the apparent accuracy of species presence/absence models and their predictions. Global Ecology and Biogeography, 20: 498-508.

Fourier, J. (1822). Thorie Analytique de la Chaleur. Didot, Paris.

Gastner, M., Newman, M. (2004). Diffusion-based method for producing density-equalizing maps. Proceedings of the National Academy of Sciences USA, 101: 7499-7504.

Gopal, S., Woodcock, C. (1994). Theory and methods for accuracy assessment of thematic maps using fuzzy sets. Photogrammetric Engineering and Remote Sensing, 60: 181-188.

Hanspach, J., Kühn, I., Schweiger, O., Pompe, S., Klotz, S. (2011). Geographical patterns in prediction errors of species distribution models. Global Ecology and Biogeography, 20: 779-788.

Haralick, R., Shanmugam, K., Dinstein, I. (1973). Textural features for image classification. IEEE Trans. Syst., Man Cybern. SMC-3, 610621.

Imre, A.R., Cseh, D., Neteler, M., Rocchini, D. (2011). Korcak dimension as a novel indicator of landscape fragmentation and re-forestation. Ecological Indicators, 11: 1134-1138.

Mandelbrot, B.B., Blumen, A. (1989). Fractal Geometry: What is it, and what does it do? Proc. R. Soc. Lond. A, 423: 3-16.

Neteler, M., Bowman, M.H., Landa, M., Metz, M. (2012). GRASS GIS: A multi-purpose open source gis. Environmental Modelling & Software, 31: 124-130.

R Development Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rényi, A., 1970. Probability Theory. North Holland Publishing Company, Amsterdam.

Ricotta, C. (2005). Additive partitioning of Rao's quadratic diversity: a hierarchical approach. Ecological Modelling, 183: 365-371.

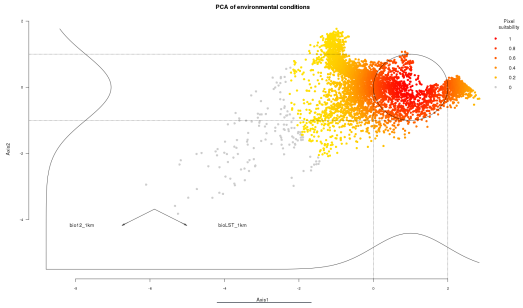
Rocchini, D., Boyd, D.S., Féret, J.B., Foody, G.M., He, K.S., Lausch, A., Nagendra, H., Wegmann, M., Pettorelli, N. (2016). Satellite remote sensing to monitor species diversity: potential and pitfalls. Remote Sensing in Ecology and Conservation, 2: 25-36.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., Chiarucci, A. (2011). Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. Progress in Physical Geography, 35: 211-226.

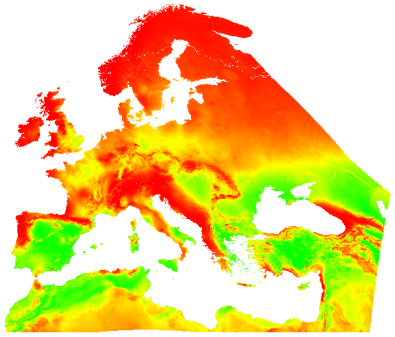
Walsh, J. (1923). A closed P set of orthonormal functions. American Journal of Mathematics, 45: 5-24.

Zadeh, L., 1965. Fuzzy sets. Information Control 8, pp. 338353.

Figures



(a) Predictor space



(b) 2D space

Figure 1: A virtual species distribution might be useful to synthesize species spread conditional to known ecological drivers. Panel a) Environmental suitability of the virtual species in the predictor space, represented with two climate variables. Panel b) The habitat suitability for the virtual species created. Suitability is represented from low in red to high in green.

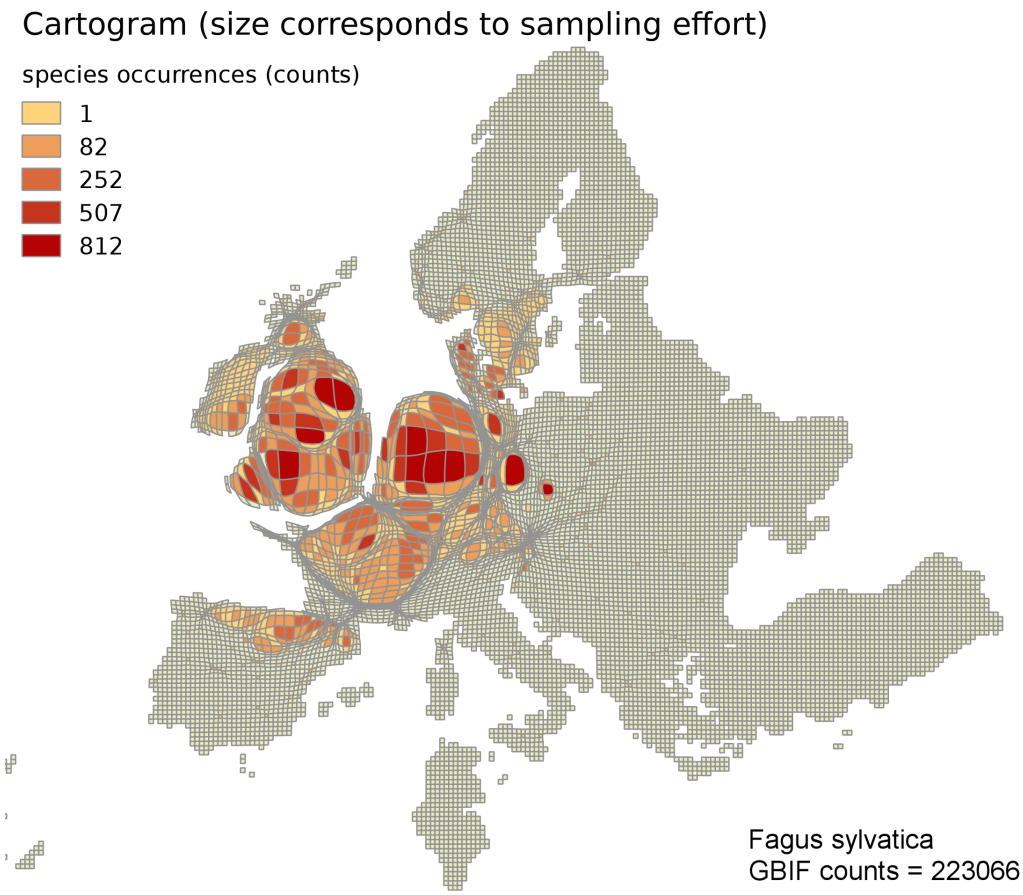


Figure 2: Cartograms can be used to show the sampling effort bias in species distribution modelling. In this case, oversampled cells are more distorted than the others; hence in such cells the higher abundance of *Fagus sylvatica* might be an artifact to to oversampling.

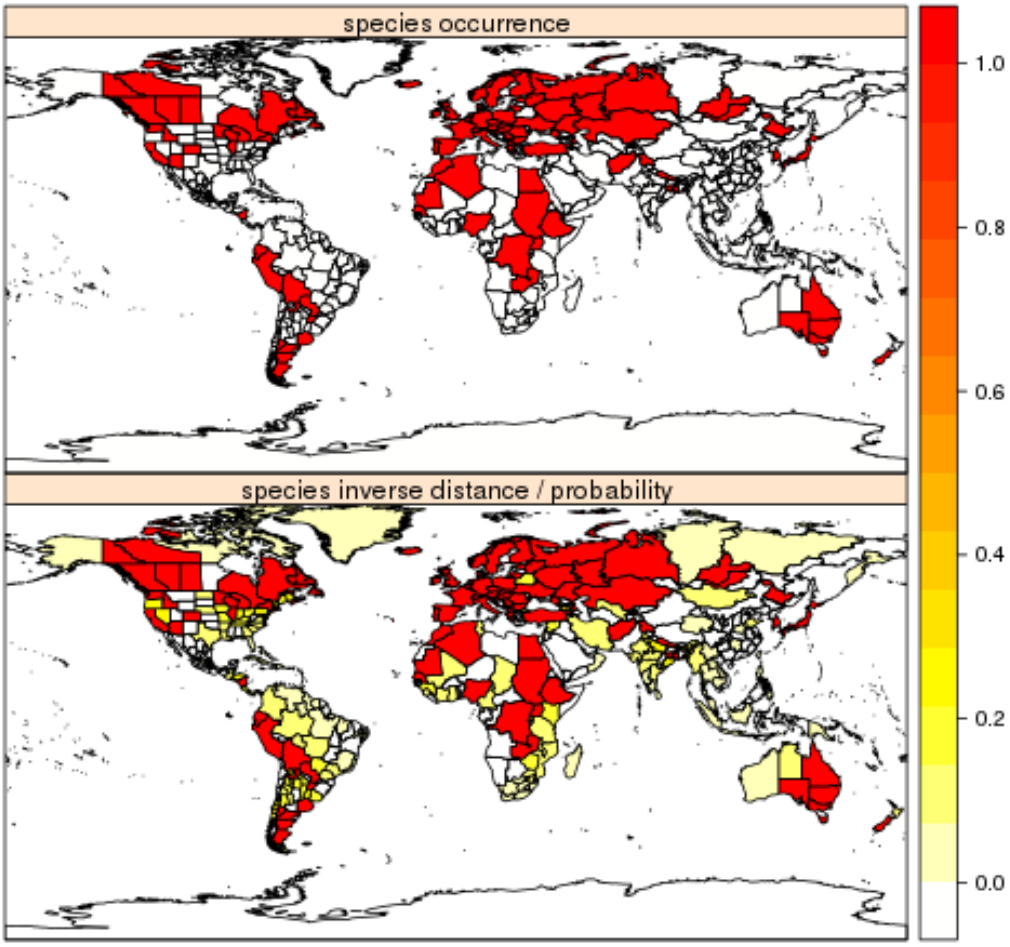


Figure 3: Representation of the presence of the foraminifera species *Keratella quadrata* and the probability (as inverse distance) of occurrence of each determined individual to that species. While the presence / absence map has obviously only red (1 - presence) and white (0 - absence) colour, the probability map based on inverse distance covers the whole range of decimal values from 0 to 1.

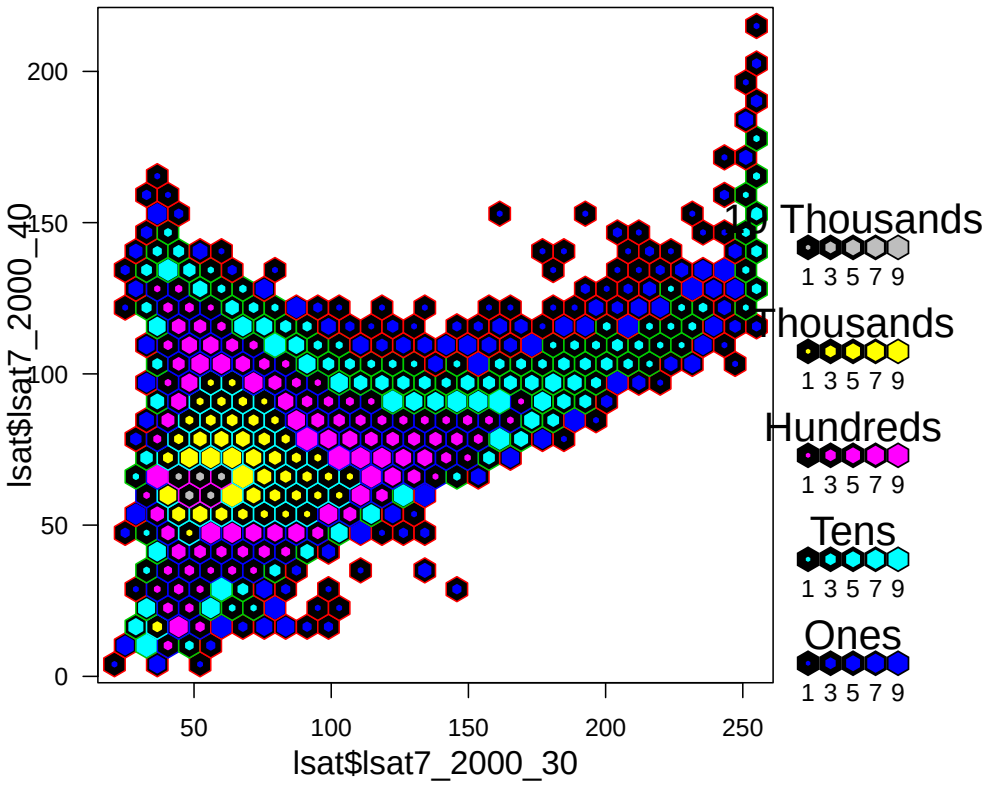
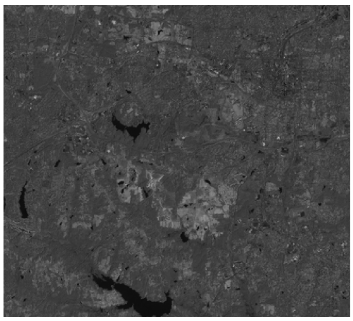
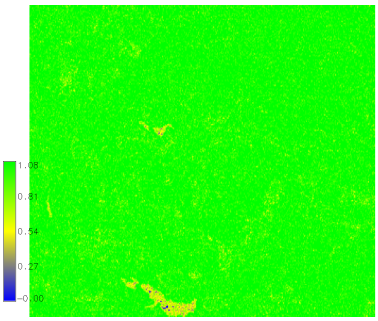


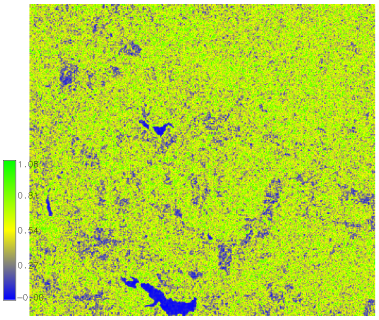
Figure 4: Starting from two Landsat ETM+ bands, hexagon binning allows to explore their relationship by also showing the amount of data per each value.



(a) Landsat ETM+ band4 NIR



(b) Entropy



(c) Variance

Figure 5: Texture measures derived from a Landsat ETM+ band.

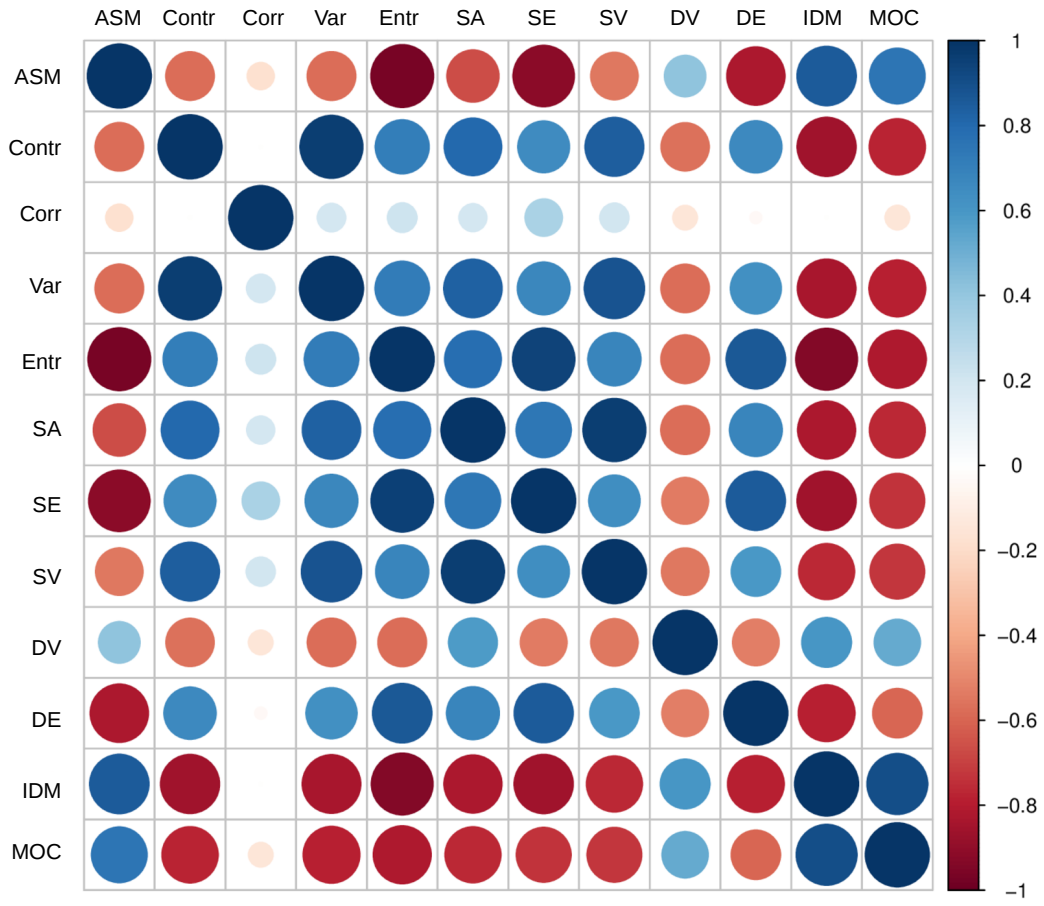


Figure 6: A corrplot by R allows to directly show the amount of correlation among RS layers. In this example, the system composed by texture measures (sensu Haralick et al. (1973)) is generally highly positively or negatively correlated. Refer to the main text for additional information on single measures' acronyms. Reproduced from **?**.

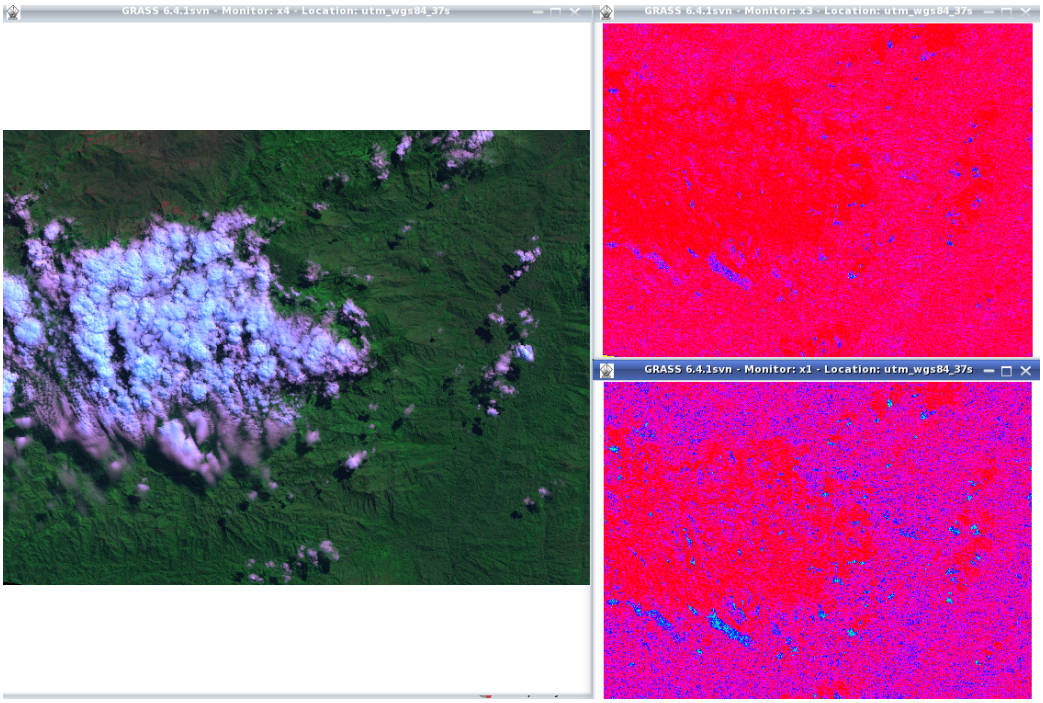


Figure 7: Starting from the same RS image (left) Rényi generalized entropy based on different alpha values can lead to different maps to better synthesize the continuous variation of ecological diversity in space. This panels are related to calculations in GRASS GIS.

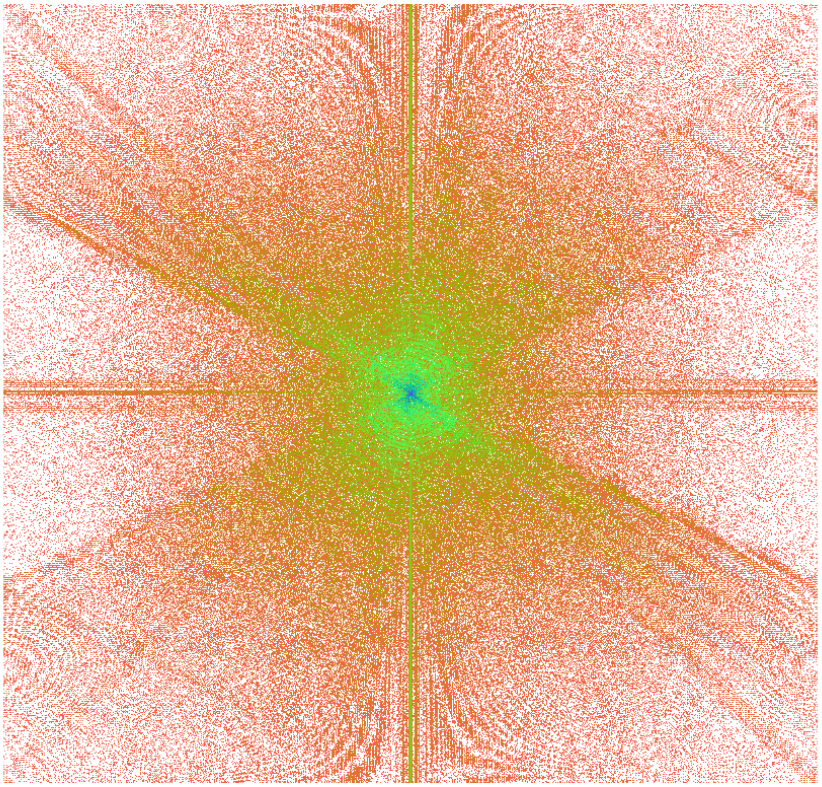


Figure 8: A Fourier image gathered applying Eq.(**??**) to a remotely sensed image. The external part of a Fourier frequency space contains high frequency values while the part near the centre contains low frequency values. Hence the higher the amount of red values (higher values) occupying the white (low values) external part, the higher will be the heterogeneity in the landscape.

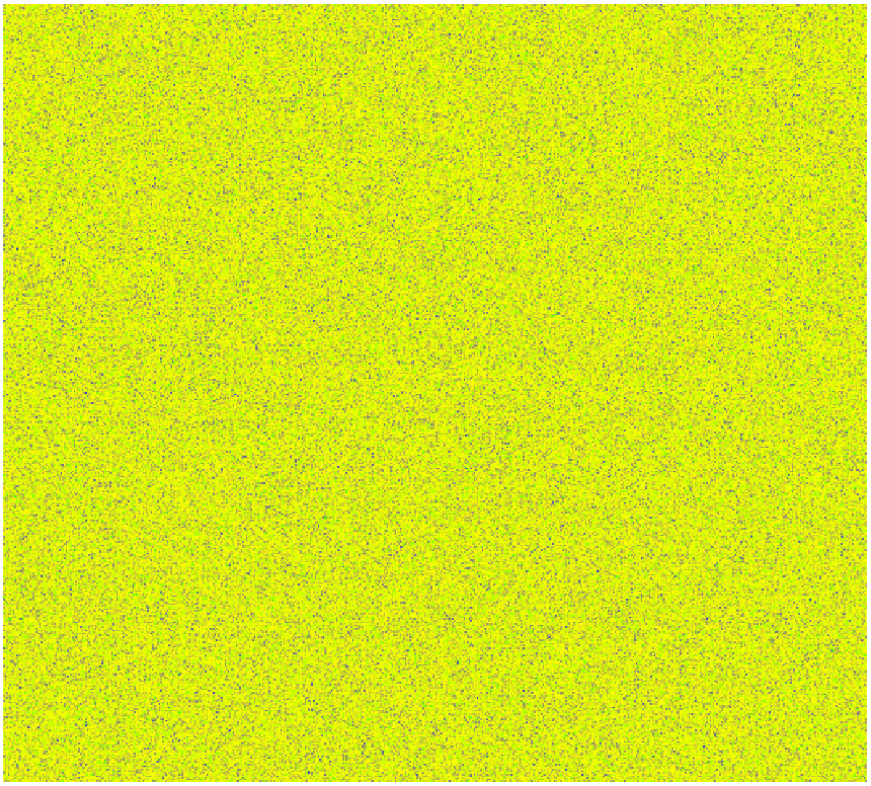
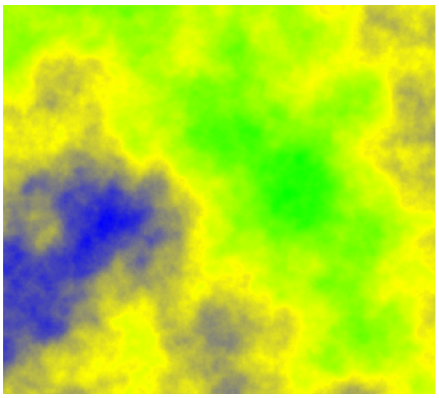
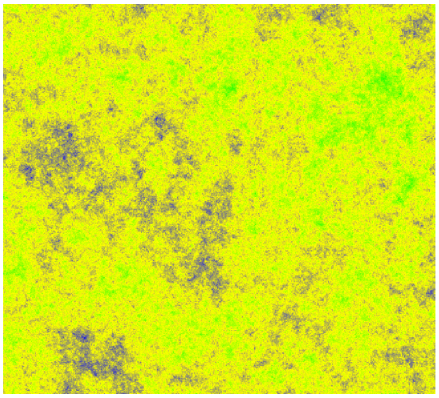


Figure 9: Random surfaces can be used to test from deviations of natural patterns from random expectations. This Figure shows a completely random surface.



(a) Fractal dimension = 2.1



(b) Fractal dimension = 2.96

Figure 10: Artificial landscapes with different fractal dimensions.